1. A sorbanállási rendszerek jellemzői

Ahhoz, hogy teljesen jellemezzünk egy sorbanállási rendszert, azonosítanunk kell azt a sztochasztikus folyamatot, amely a beérkező igényeket írja le, és meg kell adnunk a kiszolgálás szabályait és struktúráját. A beérkező folyamatot általában az egymás után beérkező igények közötti időintervallumok valószínűségeloszlása segítségével jellemezhetjük. Ezt az $A(t)$ szimbólummal jelöljük, ahol

$$A(t) = P(\text{két egymás utáni beérkezési időköz} < t).$$

A sorbanállás elméletében többnyire feltesszük, hogy az egymás utáni beérkezések közötti időközök (röviden *beérkezési időköz*), azonos eloszlású független valószínűségi változók (ezért a beérkezési folyamat ún. *felújítási folyamatot* alkot). A másik sztochasztikus mennyiség, amit meg kell adni, a beérkező igények által a csatornával szemben támasztott követelmények (munka) nagysága; ezt *kiszolgálási időnek* nevezzük és valószínűségeloszlását $B(x)$-szel jelöljük, azaz

$$B(x) = P(\text{kiszolgálási idő} < x).$$

A kiszolgálás ideje annak az időintervallumnak a hosszát jelenti, amelyet az igény a kiszolgáló egységben eltölt.

A kiszolgálás szabályára és struktúrájára vonatkozóan további mennyiségeket kell meghatározni. Ilyen jellemző a rendelkezésre álló *kiszolgálóegységek (csatornák) száma*, valamint a *befogadóképesség*, ami nem más, mint a kiszolgálóegységben és a várakozási sorban tartózkodó igények maximális száma, amit gyakran végtelennek tekintünk. A *kiszolgálási sorrend* írja le azt a szabályt, amely szerint a várakozók közül sorra kerülnek az egyes igények kiszolgálás céljából. A leggyakrabban használt kiszolgálási elvek : FIFO (First In - First Out) - érkezési sorrendben; LIFO (Last In - First Out) - fordított sorrendben történő kiszolgálások. Ha a beérkező igényeket bizonyos csoportokba tartozás szerint meg lehet különböztetni, akkor a csoportok között *prioritást* lehet megállapitani, és ezen a prioritáson alapul a kiszolgálás sorrendje. Ez az egyik legalkalmasabb ütemezési elv, mivel így az igények közötti fontossági sorrendet felállítva történik a kiszolgálás.

A prioritásos sorbanállási elvnek két fő típusa van: *abszolút* és *relatív*. Az előbbi azt jelenti, hogy ha egy igény kiszolgálása folyamatban van, és érkezik egy magasabb prioritású igény, akkor a kiszolgálás megszakad, és újra beáll a várakozási sorba. He legközelebb rákerül a kiszolgálás, akkor az kezdődhet az elejétől vagy a megszakítás helyétől. A relatív prioritásos esetben a fontosabb igény beérkezésekor a kiszolgálás nem szakad meg, hanem folytatódik, majd a befejezéskor a legfontosabb várakozó igény kiszolgálása kezdődik.

A sorbanállási rendszerek hatékonyságának és teljesítményének vizsgálatához a következő mérőszámokat fogjuk meghatározni: az *igények várakozási ideje*; a rendszerben levő *igények száma*; a *foglaltsági intervallum hossza* (vagyis az a folytonos időintervallum, amelyben a kiszolgáló egység állandóan foglalt); az *üresjárati időszakasz hossza*; a pillanatnyi *munkahátralék eloszlása*. Mindegyik mennyiség valószínűségi változó, és így teljes valószínűségszámítási jellemzésüket (vagyis eloszlásfüggvényüket) keressük, amit általában nehéz megadni, így sokszor megelégszünk az átlagos mennyiségekkel.

Az elemi sorbanállási elmélet egyrészt történeti okokból, másrészt pedig azért fontos, mert alkalmas arra, hogy szemléltesse a bonyolultabb sorbanállási rendszerek jellemzőit is.

Egyszerűség kedvéért tekintsünk először egy egykiszolgálós rendszert.

A sorbanállási rendszerek teljesítményének mérésére legalkalmasabb eszköz a torlódás vizsgálata. Legyen ϱ egy dimenzió nélküli mennyiség, amelyet a következőképpen lehet definiálni:

$$\varrho = \text{forgalmi intenzitás} = \frac{\text{átlagos kiszolgálási idő}}{\text{átlagos beérkezési időköz}}.$$

Feltételezzünk egy végtelen populációjú modellt, jelöljük a beérkezési intenzitást λ-val, ami nem más, mint az átlagos beérkezési időköz reciproka, valamint az átlagos kiszolgálási időt $1/\mu$-vel. Ekkor a következőt kapjuk:

$$\varrho = \text{érkezési intenzitás} * \text{átlagos kiszolgálási idő} = \frac{\lambda}{\mu}.$$

Az 1-nél nagyobb *forgalmi intenzitás* azt mutatja ,hogy az igények gyorsabban érkeznek, mint ahogy egy szerver (kiszolgálóegység, csatorna) ki tudná szolgálni őket. Jelölje $\chi(A)$ az A esemény karakterisztikus függvényét, azaz

$$\chi(A) = \begin{cases} 1 & \text{, ha } A \text{ teljesül,} \\ 0 & \text{, ha nem } A \text{ teljesül ,} \end{cases}$$

és $X(t) = 0$ azt az eseményt, hogy a kiszolgáló tétlen a t időpillanatban. Ekkor a szerver időegységre eső kihasználtsága

$$\frac{1}{T} \int\limits_0^T \chi\left(X(t) \neq 0\right) dt \ ,$$

ahol T egy elegendően hosszú időintervallum. Ha $T \to \infty$ esetén a fenti mennyiségeknek létezik határértéke, akkor a szerver *kihasználtságán* ezt az U_s-sel jelölt mennyiséget értjük. Továbbá 1 valószínűséggel fennáll

$$U_s = \lim_{T\to\infty} \frac{1}{T} \int\limits_0^T \chi\left((t) \neq 0\right) dt = 1 - p_0 = \frac{E\delta}{E\delta + Ei},$$

ahol p_0 annak stacionárius valószínűsége, hogy a szerver tétlen, $E\delta$ a kiszolgáló egység átlagos foglaltsági periódushosszát, Ei pedig az átlagos tétlenségi periódushosszát jelöli.

Ez az összefüggés Markov-folyamatoknál speciális esete a következő, gyakran felhasználható relációnak. Legyen $X(t)$ egy ergodikus Markov-folyamat, A pedig állapotterének egy részhalmaza. Látható, hogy $X(t)$ az idő folyamán felváltva tartózkodik A-ban és $\overline{A}$-ban. Ekkor 1 valószínűséggel

$$\lim_{T\to\infty} \frac{1}{T} \left(\int_0^T \chi(X(t) \in A) dt \right) = \sum_{i\in A} p_i$$

$$= \frac{m(A)}{m(A) + m(\overline{A})},$$

ahol $m(A)$ és $m(\overline{A})$ az A ill. az $\overline{A}$ részhalmazban való átlagos tartózkodási időt jelöli egy ciklus alkalmával, p_i pedig az $X(t)$ folyamat ergodikus eloszlása.

Egy m párhuzamos szerverből álló rendszerben T idő alatt átlagosan $\lambda T/m$ igény érkezik szerverenként, feltéve, hogy a forgalom egyenletes eloszlású az m kiszolgáló egység között. Ha minden beérkezett kérés kiszolgálása átlagosan $1/\mu$ ideig tart, akkor a szerver teljes foglaltsági idejének várható értéke $\lambda T/m\mu$. Osszuk el ezt a mennyiséget T-vel, így

$$\varrho = \frac{\lambda}{m\mu}.$$

Mivel a kihasználtság maximum 1 lehet, így az m szerveres rendszer kihasználtsági tényezőre vonatkozó korrekt kifejezés:

$$\varrho = \min\left\{\frac{\lambda}{m\mu}, 1\right\}.$$

Másik gyakran használt teljesítménymérő eszköz a *rendszer átbocsátóképességének* vizsgálata. Ezt a mennyiséget úgy definiálhatjuk, mint az időegységenként kiszolgált igények átlagos számát. m szerveres rendszerben minden időegység alatt $m\varrho\mu$ igény kiszolgálása fejeződik be, így az

$$\text{átbocsátóképesség} = m\varrho\mu = \min\{\lambda, m\mu\}.$$

Ami azt jelenti, hogy az átbocsátóképesség ekvivalens a λ érkezési intenzitással, amennyiben a λ kisebb, mint a maximális kiszolgálási sebesség ($m\mu$), azon túl az átbocsátóképesség beáll $m\mu$-re.

Az igények szempontjából a legjelentősebb teljesítménymérő eszköz az az idő, amit a várakozási sorban vagy a rendszerben töltenek. Definiáljuk a W_j *várakozási időt*, mint a j-dik igény várakozási sorban eltöltött idejét, és a T_j *válaszidőt*, mint az igény által e rendszerben eltöltött teljes időt. Ezen jelöléseket használva a következő egyenlőséget kapjuk:

$$T_j = W_j + S_j,$$

ahol S_j a kiszolgálási időt jelöli. W_j és T_j is valószínűségi változó, várható értékük $\overline{W_j}$ és $\overline{T_j}$ alkalmas a rendszer teljesítményének mérésére.

A rendszer teljesítményének vizsgálata történhet a *várakozási sor hosszának* mérésével is. A $Q(t)$ valószínűségi változó jelentse a t időpillanatban a sorban található igények számát, és $X(t)$ a t időpillanatban a *rendszerben található igények számát.* Egy rendszerben levő igény vagy a várakozási sorban van, vagy éppen kiszolgálás alatt áll, tehát m szerveres rendszer esetén:

$$Q(t) = \max\{0, X(t) - m\}.$$

Mielőtt rátérnénk az elemi sorbanállási rendszerek vizsgálatára, néhány, Kendalltól származó jelölést vezetünk be, melyek segítségével osztályozhatjuk őket:

A/B/m/K/N

ahol

A: a beérkezési időközök eloszlásfüggvénye,
B: a kiszolgálási idő eloszlásfüggvénye,
m: a kiszolgálók száma,
K: a rendszer befogadóképessége, azaz a kiszolgálóegységben és a várakozási sorban tartózkodó igények maximális száma,
N: az igényforrás számossága.

Ha az említett eloszlások exponenciálisak, akkor az M jelölést használjuk. Továbbá, ha a befogadóképesség vagy az igényforrás számossága végtelen, akkor ezeket a jelöléseket elhagyjuk.

Így pl. az M/M/1 rendszer, egy egy kiszolgálós Poisson beérkezéssel és exponenciális kiszolgálási idővel jellemzett rendszert jelöl. Az M/G/m rendszernél a beérkezések Poisson-folyamat szerint történnek, a kiszolgálási idők általános eloszlásúak, és m szerver áll rendelkezésünkre. Az M/M/r/N/N rendszer esetén az igények egy N elemű forrásból származnak ahol exponenciális eloszlású ideig tartózkodnak, a kiszolgálást r egység végzi exponenciális eloszlású ideig.